Robust High Dimensional Sparse Regression and Matching Pursuit

Yudong Chen, Constantine Caramanis and Shie Mannor

Abstract

In this paper we consider high dimensional sparse regression, and develop strategies able to deal with arbitrary – possibly, severe or coordinated – errors in the covariance matrix X. These may come from corrupted data, persistent experimental errors, or malicious respondents in surveys/recommender systems, etc. Such non-stochastic error-in-variables problems are notoriously difficult to treat, and as we demonstrate, the problem is particularly pronounced in high-dimensional settings where the primary goal is *support recovery* of the sparse regressor. We develop algorithms for support recovery in sparse regression, when some number n_1 out of $n + n_1$ total covariate/response pairs are *arbitrarily (possibly maliciously) corrupted*. We are interested in understanding how many outliers, n_1, we can tolerate, while identifying the correct support. To the best of our knowledge, neither standard outlier rejection techniques, nor recently developed robust regression algorithms (that focus only on corrupted response variables), nor recent algorithms for dealing with stochastic noise or erasures, can provide guarantees on support recovery. Perhaps surprisingly, we also show that the natural brute force algorithm that searches over all subsets of n covariate/response pairs, and all subsets of possible support coordinates in order to minimize regression error, is remarkably poor, unable to correctly identify the support with even $n_1 = O(n/k)$ corrupted points, where k is the sparsity. This is true even in the basic setting we consider, where all authentic measurements and noise are independent and sub-Gaussian. In this setting, we provide a simple algorithm – no more computationally taxing than OMP – that gives stronger performance guarantees, recovering the support with up to $n_1 = O(n/(\sqrt{k} \log p))$ corrupted points, where p is the dimension of the signal to be recovered.

I. Introduction

Linear regression and sparse linear regression seek to express a response variable as the linear combination of (a small number of) covariates. They form one of the most basic procedures in statistics, engineering, and science. More recently, regression has found increasing applications in the high-dimensional regime, where the number of variables, p, is much larger than the number of measurements or observations, n. Applications in biology, genetics, as well as in social networks, human behavior prediction and recommendation, abound, to name just a few. The key structural property exploited in high-dimensional regression, is that the regressor is often sparse, or near sparse, and as much recent research has demonstrated, in many cases it can be efficiently recovered, despite the grossly underdetermined nature of the problem (e.g., [8], [6], [4], [12], [31]). Another common theme in large-scale learning problems – particularly problems in the high-dimensional regime – is that we not only have big data, but we have dirty data. Recently, attention has focused on the setting where the output (or response) variable and the matrix of covariates are plagued by erasures, and/or by stochastic additive noise [23], [26], [27], [9], [10]. Yet many applications, including those mentioned, may suffer from persistent errors, that are ill-modeled by stochastic distribution; indeed, many applications, particularly those modeling human behavior, may exhibit maliciously corrupted data.

This paper is about extending the power of regression, and in particular, sparse high-dimensional regression, to be robust to this type of noise. We call this *deterministic* or *cardinality constrained* robustness, because rather than restricting the magnitude of the noise, or any other such property of the noise, we merely assume there is a bound on how many data points, or how many coordinates of every single covariate, are corrupted. Other than this number, we make absolutely no assumptions on what the adversary can do – the adversary is virtually unlimited in computational power and knowledge about our algorithm and about the authentic points. There are two basic models we consider. In both, we

ydchen@utexas.edu, constantine@utexas.edu, shie@ee.technion.ac.il

assume there is an underlying generative model: $y = X\beta^* + e$, where X is the matrix of covariates, e is sub-Gaussian noise. In the *row-corruption model*, we assume that each pair of covariates and response we see is either due to the generative model, i.e., (y_i, X_i), or is corrupted in some arbitrary way, with the only restriction that at most n_1 such pairs are corrupted. In the *distributed corruption model*, we assume that y and *each column* of X, has n_1 elements that are arbitrarily corrupted (evidently, the second model is a strictly harsher corruption model). Building efficient regression algorithms that recover at least the support of β^* accurately subject to even such deterministic data corruption, greatly expands the scope of problems where regression can be productively applied. The basic question is when is this possible – how big can n_1 be, while still allowing correct recovery of the support of β^*.

Many sparse-regression algorithms have been proposed, and their properties under clean observations are well understood; we survey some of these results in the next two sections. Also well-known, is that the performance of standard algorithms (e.g., Lasso, Orthogonal Matching Pursuit) breaks down even in the face of just a few corrupted points or covariate coefficients. As more work has focused on robustness in the high-dimensional regime, it has also become clear that the techniques of classical robust statistics such as outlier removal preprocessing steps cannot be applied to the high-dimensional regime [13], [18]. The reason for this lies in the high dimensionality. In this setting, identifying outliers *a priori* is typically impossible: outliers might not exhibit any strangeness in the ambient space due to the high-dimensional noise (see [33] for a further detailed discussion), and thus can be identified only when the true low-dimensional structure is (at least approximately) known; on the other hand, the true structure cannot be computed by ignoring outliers. Other classical approaches have involved replacing the standard mean squared loss with a trimmed variant or even median squared loss [16]; first, these are non convex, and second, it is not clear that they provide any performance guarantees, especially in high dimensions.

Recently, the works in [20], [25], [32], [22] have proposed an approach to handle arbitrary corruption in the *response variable*. As we show, this approach faces serious difficulties when the *covariates* is also corrupted, and is bound to fail in this setting. One might modify this approach in the spirit of Total Least Squares (TLS) [36] to account for noise in the covariates (discussed in Section III), but it leads to highly non convex problems. Moreover, the approaches proposed in these papers are the natural convexification of the (exponential time) brute force algorithm that searches over all subsets of covariate/response pairs (i.e., rows of the measurement matrix and corresponding entries of the response vector) and subsets of the support (i.e., columns of the measurement matrix) and then returns the vector that minimizes the regression error over the best selection of such subsets. Perhaps surprisingly, we show that the brute force algorithm itself has remarkably weak performance. Another line of work has developed approaches to handle stochastic noise or small bounded noise in the covariates [17], [26], [27], [23], [9]. The corruption models in there, however, are different from ours which allows arbitrary and malicious noise; those results seem to depend crucially on the assumed structure of the noise and cannot handle the setting in this paper.

More generally, even beyond regression, in, e.g., robust PCA and robust matrix completion [7], [5], [34], [11], [21], recent robust recovery in high dimensions results have for the most part depended on convex optimization formulations. We show in Section IV that for our setting, *convex-optimization based approaches that try to relax the brute-force formulation fail to recover support, with even a constant number of outliers*. Accordingly, we develop a different line of robust algorithms, focusing on Greedy-type approaches like Matching Pursuit (MP).

In summary, to the best of our knowledge, no robust sparse regression algorithm has been proposed that can provide performance guarantees, and in particular, guaranteed support recovery, under *arbitrarily and maliciously corrupted* covariates and response variables.

We believe robustness is of great interest both in practice and in theory. Modern applications often involve "big but dirty data", where outliers are ubiquitous either due to adversarial manipulation or to the fact some samples are generated from a model different from the assumed one. It is thus desirable to develop robust sparse regression procedures. From a theoretical perspective, it is somewhat surprising that the addition of a few outliers can transform a simple problem to a hard one; we discuss the difficulties in more detail in the subsequent sections.

Paper Contributions: In this paper, we propose and discuss a simple (in particular, efficient) algorithm for robust sparse regression, for the setting where both covariates and response variables are arbitrarily corrupted, and show that our algorithm guarantees support recovery under far more corrupted covariate/response pairs than any other algorithm we are aware of. We briefly summarize our contributions here:

1) We consider the corruption model where n_1 rows of the covariate matrix X and the response vector y are arbitrarily corrupted. We demonstrate that other algorithms we are aware of, including standard convex optimization approaches and the natural brute force algorithm, have very weak (if any) guarantees for support recovery.
2) For the corruption model above, we give support recovery guarantees for our algorithm, showing that we correctly recover the support with $n_1 = O(n/(\sqrt{k} \log p))$ arbitrarily corrupted response/covariate pairs.
3) We consider a stronger corruption model, where instead of n_1 corrupted rows of the matrix X, *each column can have up to n_1 arbitrarily corrupted entries*. We show that our algorithm also works in this setting, with precisely the same recovery guarantees. To the best of our knowledge, this problem has not been previously considered.

II. PROBLEM SETUP

We consider the problem of sparse linear regression. The unknown parameter $\beta^* \in \mathbb{R}^p$ is assumed to be k-sparse ($k < p$), i.e., has only k nonzeros. The observations take the form of covariate-response pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n + n_1$. Among these, n pairs are authentic samples obeying the following linear model

$$y_i = \langle x_i, \beta^* \rangle + e_i,$$

where e_i is additive noise, and $p \geq n$. For corruption, we consider the following two models.

Definition 1 (Row Corruption). The n_1 pairs are arbitrarily corrupted, with *both* x_i and y_i being potentially corrupted.

Definition 2 (Distributed Corruption). We allow arbitrary corruption of *any $n_1/2$ elements of each column* of the covariate matrix X and of the response y.

In particular, the corrupted entries need not lie in the same n_1 rows. Clearly this includes the previous model as a special case up to a constant factor of 2.

Note that in both models, we impose *no assumption whatsoever on the corrupted pairs*. They might be unbounded, non-stochastic, and even dependent on the authentic samples. They are unconstrained other than in their cardinality – the number of rows or coefficients corrupted. We illustrate both of these corruption models pictorially in Figure 1.

Goal: Given these observations $\{(x_i, y_i)\}$, the goal is to obtain a reliable estimate $\hat{\beta}$ of β^* with correct support and bounded error $\left\| \hat{\beta} - \beta^* \right\|_2$. A fundamental question, therefore, is to understand in each given model, given p, n, and k, how many outliers (n_1) an estimator can handle.

Remark 3. This is a strong notion of robustness. In particular, requiring support recovery is a more stringent requirement than requiring, for example, bounded distance to the true solution, or bounded loss degradation. It is also worth noting that robustness here is a completely different notion than robustness in the sense of Robust Optimization (e.g., [2]). There, we seek a solution that minimizes the error we incur when an adversary perturbs our loss function. In contrast, here, here is a true generative model that obeys the structural assumptions of the problem (namely, sparsity of β^*), but an adversary corrupts the data that ordinarily provide us with the true input-output behavior of that model.

We emphasize that our setting is fundamentally different from those that only allow corruptions in y, and robust sparse regression techniques that only consider corruption in y are bound to fail in our setting.

We elaborate on this point in Section IV and V. Moreover, under the distributed corruption model, there is no hope of considering an equivalent model with only corruptions in y: all entries of y could be corrupted – an absurd setting to hope for a solution.
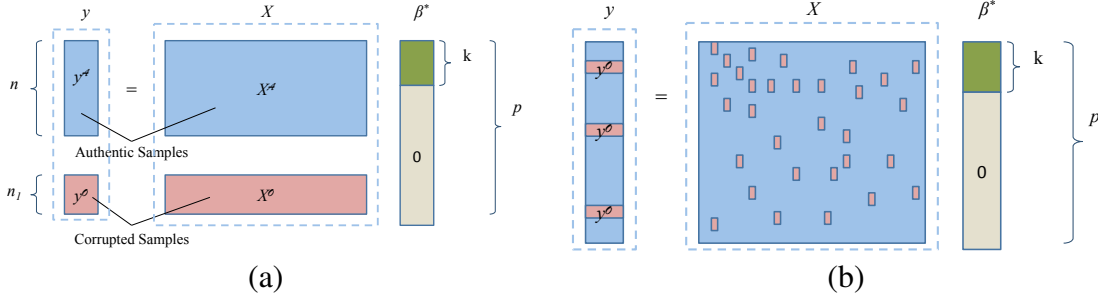

Figure 1: Consider identifying the dominant genetic markers for a disease; entire experiments might fail, or different gene expression levels might be erroneously read across different experiments. Panel (a) corresponds to the first type of error, where n_1 samples (y_i, X_i) are arbitrarily corrupted. This is our first error model. Panel (b) illustrates an instance of the second scenario, and thus our second error model: in each experiment, several covariates may be misread, but no single covariate is misread more than n_1 times.

III. RELATED WORK

Under the high-dimensional setting $p \geq n$, there is a large body of literature on sparse recovery when there is no corruption. It is now well-known that recovery of β^* is possible only when the covariate matrix X satisfies certain conditions, such as the Restricted Isometry/Eigenvalue Property [6], [3], Mutual Incoherence Condition [14] or Exact Recovery Condition [29]. Various ensembles of random matrices are shown to satisfy these conditions with high probability. Many estimators have been proposed, most notably Basis Pursuit (a.k.a. Lasso) [28], [14], which solves an ℓ_1-regularized least squares problem

$$\min_{\beta} \quad \|y - X\beta\|_2 + \lambda \|\beta\|_1 ,$$

as well as Orthogonal Matching Pursuit (OMP) [14], [29], which is a greedy algorithm that estimates the support of β^* sequentially. Both Lasso and OMP, as well as many other estimators, are guaranteed to recover β^* with good accuracy when X is well-conditioned, and the number of observations satisfies $n \gtrsim k \log p$. (Here we mean there exists a constant c, independent of k, n, p, such that the statement holds. We use this notation throughout the paper.) Moreover, this condition is also shown to be necessary; see, e.g., [30].

Most existing methods are not robust to outliers; for example, standard Lasso and OMP fail even if only one entry of X or y is corrupted. One might consider a natural modification of Lasso in the spirit of Total Least Squares, and solve

$$\min_{\beta, E} \|(X - E)\beta - y\|_2 + \lambda\|\beta\|_1 + \eta\|E\|_*, \tag{1}$$

where E accounts for corruption in the covariate matrix, and $\|\cdot\|_*$ is a norm. When E is known to be row sparse (as is the case in our row-corruption model), one might choose $\|\cdot\|_*$ to be $\|\cdot\|_{1,2}$ or $\|\cdot\|_{1,\infty}$[1]; the work in [36] considers using $\|\cdot\|_* = \|\cdot\|_F$ (similar to TLS), which is more suitable when E is dense yet bounded. The optimization problem (1) is, however, highly non convex due to the bilinear term $E\beta$, and no tractable algorithm with provable performance guarantees is known.

---

[1] $\|E\|_{1,2}$ ($\|E\|_{1,\infty}$) is the sum of the ℓ^2 (ℓ^∞, respectively) norms of the rows of E.

Another modification of Lasso accounts for the corruption in the response via an additional variable z [20], [25], [32], [22]:

$$\min_{\beta,z} \|X\beta - y - z\|_2 + \lambda \|\beta\|_1 + \gamma \|z\|_1. \tag{2}$$

We call this approach Justice Pursuit (JP) after [20]. Unlike the previous approach, the problem (2) is convex. In fact, it is the natural convexification of the brute force algorithm:

$$\begin{aligned} \min_{\beta,z} \quad & \|X\beta - y - z\|_2 \\ \text{s.t.} : \quad & \|\beta\|_0 \le k \\ & \|z\|_0 \le n_1, \end{aligned} \tag{3}$$

where $\|u\|_0$ denotes the number of nonzero entries in u. It is easy to see (and well known) that the so-called Justice Pursuit relaxation (2) is equivalent to minimizing the Huber loss function plus the ℓ_1 regularizer, with an explicit relation between γ and the parameter of the Huber loss function [15]. Formulation (2) has excellent recovery guarantees when *only* the response variable is corrupted, delivering exact recovery under a constant fraction of outliers. However, we show in the next section that a broad class of convex optimization-based approaches, with (2) as a special case, fail when the covariate X is also corrupted. In the subsequent section, we show that even the original brute force formulation is problematic: while it can recover from some number n_1 of corrupted rows, that number is order-wise worse than what the algorithm we give can guarantee.

We also note that neither the brute force algorithm above, nor its relaxation, JP, are appropriate for our second model for corruption. Indeed, in this setting, modeling via JP would require handling the setting where *every single entry* of the output variable, y, is corrupted, something which certainly cannot be done.

For standard linear regression problems in the classical scaling $n \gg p$, various robust estimators have been proposed, including M-, R-, and S-estimators [18], [24], as well as those based on ℓ_1-minimization [19]. Many of these estimators lead to non-convex optimization problems, and even for those that are convex, it is unclear how they can be used in the high-dimensional scaling with sparse β^*. Another difficulty in applying classical robust methods to our problems arises from the fact that the covariates, x_i, also lie in a high-dimensional space, and thus defeat many outlier detection/rejection techniques that might otherwise work well in low-dimensions. Again, for our second model of corruption, outlier detection seems even more hopeless.

IV. Failure of the Convex Optimization Approach

We consider a broad class of convex optimization-based approaches of the following form:

$$\begin{aligned} \min_{\beta} \quad & f(y - X\beta) \\ \text{s.t.} \quad & h(\beta) \le R. \end{aligned} \tag{4}$$

Here R is a radius parameter that can be tuned. Both $f(\cdot)$ and $h(\cdot)$ are convex functions, which can be interpreted as a loss function (of the residual) and a regularizer (of β), respectively. For example, one may take $f(v) = \min_z \|v - z\|_2 + \gamma \|z\|_1$ and $h(\beta) = \|\beta\|_1$, which recovers the Justice Pursuit (2) by Lagrangian duality; note that this $f(v)$ is convex because

$$\|\alpha v_1 + (1-\alpha)v_2 - z\|_2 + \gamma\|z\|_1 \le \alpha(\|v_1 - z\|_2 + \gamma\|z\|_1) + (1-\alpha)(\|v_1 - z\|_2 + \gamma\|z\|_1),$$

by sub-additivity of norms. The function $f(\cdot)$ can also be any other robust convex loss function including the Huber loss function.

We assume that $f(\cdot)$ and $h(\cdot)$ obey a very mild condition, which is satisfied by any non-trivial loss function and regularizer that we know of. In the sequel we use $[z_1; z_2]$ to denote the concatenation of two column vectors z_1 and z_2.

Definition 4 (Standard Convex Optimization (SCO) Condition). We say $f(\cdot)$ and $h(\cdot)$ satisfy the SCO Condition if $\lim_{\alpha \to \infty} f(\alpha v) = \infty$ for all $v \neq 0$, $f([v_1; v_2]) \geq f([0; v_2])$ for all v_1, v_2, and $h(\cdot)$ is invariant under permutation of coordinates.

We also assume $R \geq h(\beta^*)$ because otherwise the formulation is not consistent even when there are no outliers. The following theorem shows that under this assumption, the convex optimization approach fails when both X and y are corrupted. We only show this for our first corruption model, since it is a special case of the second distributed model. As illustrated in Figure 1, let $\mathcal{A}$ and $\mathcal{O}$ be the (unknown) sets of indices corresponding to authentic and corrupted observations, respectively, and $X^{\mathcal{A}}$ and $X^{\mathcal{O}}$ be the authentic and corrupted rows of the covariate matrix $X = [x_1, \ldots, x_{n+n_1}]^\top$. The vectors $y^{\mathcal{A}}$ and $y^{\mathcal{O}}$ are defined similarly. Also let Λ^* be the support of β^*. With this notation, we have the following.

Theorem 5. *Suppose f and h satisfy the SCO Condition. When $n_1 \geq 1$ and $k \geq 1$, the adversary can corrupt X and y in such a way that for all R with $R \geq h(\beta^*)$ the optimal solution does not have the correct support.*

Proof: Recall that $y = [y^{\mathcal{A}}; y^{\mathcal{O}}]$ and $X = [X^{\mathcal{A}}; X^{\mathcal{O}}]$ with $y^{\mathcal{A}} = X^{\mathcal{A}}\beta^* + e$, and Λ^* is the true support. The adversary fixes some set $\hat{\Lambda}$ disjoint from the true support Λ^* with $|\hat{\Lambda}| = |\Lambda^*|$. It then chooses $\hat{\beta}$ and $y^{\mathcal{O}}$ such that $\hat{\beta}_{\hat{\Lambda}} = \beta^*_{\Lambda^*}$, $\hat{\beta}_{\hat{\Lambda}^c} = 0$, and $y^{\mathcal{O}} = X^{\mathcal{O}}\hat{\beta}$ with $X^{\mathcal{O}}$ to be determined later. By assumption we have $h(\hat{\beta}) = h(\beta^*) \leq R$, so $\hat{\beta}$ is feasible. Its objective value is $f(y - X\hat{\beta}) = f([y^{\mathcal{A}} - X^{\mathcal{A}}_{\hat{\Lambda}}\beta^*_{\Lambda^*}; 0]) \leq C$ for some finite constant C. The adversary further chooses $X^{\mathcal{O}}$ such that $X^{\mathcal{O}}_{\Lambda^*} = 0$ and $X^{\mathcal{O}}_{\hat{\Lambda}}$ is large. Any $\tilde{\beta}$ supported on Λ^* has objective value

$$f(y - X\tilde{\beta}) = f([y^{\mathcal{A}} - X^{\mathcal{A}}\tilde{\beta}; X^{\mathcal{O}}(\hat{\beta} - \tilde{\beta})]) = f([y^{\mathcal{A}} - X^{\mathcal{A}}\tilde{\beta}; X^{\mathcal{O}}_{\hat{\Lambda}}\beta^*_{\Lambda^*}]) \geq f([0; X^{\mathcal{O}}_{\hat{\Lambda}}\beta^*_{\Lambda^*}]),$$

which can be made bigger than C under the SCO Condition. Therefore, any solution $\tilde{\beta}$ with the correct support Λ^* has a higher objective value than $\hat{\beta}$, and thus is not the optimal solution. ∎

Our proof proceeds by using a simple corruption strategy. Certainly, there are natural approaches to deal with this specific example, e.g., removing entries of X with large values. But discarding such large-value entries is not enough, as there may exist more sophisticated corruption schemes where simple magnitude-based clipping is ineffective. We illustrate this with a concrete example in the simulation section, where Justice Pursuit along with large-value-trimming fails to recover the correct support. Indeed, this example serves merely to illustrate more generally the inadequacy of a purely convex-optimization-based approach.

More importantly, while the idea of considering an unbounded outlier is not new and has been used in classical Robust Statistics and more recently in [35], the above theorem highlights the sharp contrast between the success of convex optimization (e.g., JP) under corruption in only y, and its complete failure when both X and y are corrupted. Corruptions in X not only break the linear relationship between y and X, but also destroy properties of X necessary for existing sparse regression approaches. In the high dimensional setting where support recovery is concerned, there is a fundamental difference between the hardness of the two corruption models.

V. THE NATURAL BRUTE FORCE ALGORITHM

The brute force algorithm (3) can be restated as follows: it looks at all possible $n \times k$ submatrices of X and picks the one that gives the smallest regression error w.r.t. the corresponding subvector of y. Formally, let $X^{\mathcal{S}}_\Lambda$ denote the submatrix of X corresponding to row indices $\mathcal{S}$ and column indices Λ, and let $y^{\mathcal{S}}$ denote the subvector of y corresponding to indices $\mathcal{S}$. The algorithm solves

$$\min_{\theta \in \mathbb{R}^k, \mathcal{S}, \Lambda} \quad \left\| y^{\mathcal{S}} - X^{\mathcal{S}}_\Lambda \theta \right\|_2 \tag{5}$$
$$s.t. \quad |\mathcal{S}| = n,$$
$$|\Lambda| = k.$$

Suppose the optimal solution is $\hat{\mathcal{S}}, \hat{\Lambda}, \hat{\theta}$. Then, the algorithm outputs $\hat{\beta}$ with $\hat{\beta}_{\Lambda} = \theta$ and $\hat{\beta}_{\Lambda^c} = 0$. Note that this algorithm has exponential complexity in n and k, and $\mathcal{S}^c$ can be considered as an operational definition of outliers. We show that even this algorithm has poor performance and cannot handle large n_1.

To this end, we consider the simple Gaussian design model, where the entries of $X^{\mathcal{A}}$ and e are independent zero-mean Gaussian random variables with variance $\frac{1}{n}$ and $\frac{\sigma_e}{n}$, respectively. The $\frac{1}{n}$ factor is simply for normalization and no generality is lost. We consider the setting where $\sigma_e^2 = k$ and $\beta_{\Lambda^*}^* = [1, \ldots, 1]^\top$. If $n_1 = 0$, existing methods (e.g., Lasso and standard OMP), and the brute force algorithm as well, can recover the support of β^* with high probability provided $n \gtrsim k \log p$. Here and henceforth, by *with high probability (w.h.p.)* we mean with probability at least $1 - p^{-2}$. However, when there are outliers, we have the following negative result.

Theorem 6. *Under the above setting, if $n \gtrsim k^3 \log p$ and $n_1 \gtrsim \frac{3n}{k+1}$, then the adversary can corrupt X and y in such a way that the brute force algorithm does not output the correct support Λ^*.*

The proof is given in Section IX. We believe the condition $n \gtrsim k^3 \log p$ is an artifact of our proof and is not necessary. This theorem shows that the brute force algorithm can only handle $O\left(\frac{n}{k}\right)$ outliers. In the next section, we propose a simple, tractable algorithm that *outperforms* this brute force algorithm and can handle $O\left(\frac{n}{\sqrt{k}}\right)$ outliers.

VI. PROPOSED APPROACH: ROBUST MATCHING PURSUIT

The discussion in the last two sections demonstrates that standard techniques for high-dimensional statistics and robust statistics are inadequate to handle our problem. As mentioned in the introduction, we believe the key to obtaining an effective robust estimator for high-dimensional data, is simultaneous structure identification and outlier rejection. In particular, for the sparse recovery problem where the observations $Z_i = (x_i, y_i)$ reside in a high-dimensional space, it is crucial to utilize the low-dimensional structure of β^* and perform outlier rejection in the "right" low-dimensional space in which β^* lies. In this section, we propose a candidate algorithm, called Robust Matching Pursuit (RoMP), which is based on this intuition.

Standard MP estimates the support of β^* sequentially. At each step, it selects the column of X which has the largest (in absolute value) inner product with the current residual r, and adds this column to the set of previously selected columns. The algorithm iterates until some stopping criterion is met. If the sparsity level k of β^* is known, then one may stop MP after k iterations.

To successfully recover the support of β^*, standard MP relies on the fact that for well-conditioned X, the inner product $h(j) = \langle r, X_j \rangle$ is close to β_j, and thus a large value of $h(j)$ indicates a nonzero β_j. When outliers are present, MP fails because the $h(j)$'s may be distorted significantly by maliciously corrupted x_i's and y_i's. To protect against outliers, it is crucial to obtain a robust estimate of $h(j)$. This motivates our robust version of MP.

The proposed Robust Matching Pursuit algorithm (RoMP) is summarized in Algorithms 1 and 2. Similar to standard MP, it selects the columns of X with highest inner products with the residual. There are two main differences from standard MP. The key difference is that we compute a robust version of inner product by trimming large points. Also, there is no iterative procedure – we take the inner products between all the columns of X and the response vector y and selects the top k ones; this leads to a simpler analysis.

The key idea behind RoMP is that it effectively reduces a high-dimensional robust regression problem to a much easier low-dimensional (2-D) problem, one that is induced by the sparse structure of β^*. Outlier rejection is performed in this low-dimensional space, and the support of β^* is estimated along the way. Hence, this procedure fulfils our previous intuition of simultaneous structure identification and outlier rejection.

Our algorithm requires two parameters, n_1 and k. We discuss how to choose these parameters after we present the performance guarantees in the next section.

Algorithm 1 Robust Matching Pursuit (RoMP)

Input: X, y, k, n_1.

For $j = 1, \ldots, p$, compute the trimmed inner product (see Algorithm 2):

$$h(j) = \text{trimmed-inner-product}(y, X_j, n_1).$$

Sort $\{|h(j)|\}$ and select the k largest ones.
Let $\hat{\Lambda}$ be the set of selected indices.
Set $\hat{\beta}_j = h(j)$ for $j \in \hat{\Lambda}$ and 0 otherwise.
Output: $\hat{\beta}$

Algorithm 2 Trimmed Inner Product

Input: $a \in \mathbb{R}^N$, $b \in \mathbb{R}^N$, n_1
Compute $q_i = a_i b_i$, $i = 1, \ldots, N$.
Sort $\{|q_i|\}$ and select the smallest $(N - n_1)$ ones.
Let Ω be the set of selected indices.
Output: $h = \sum_{i \in \Omega} p_i$.

VII. PERFORMANCE GUARANTEES FOR RoMP

We are interested in finding conditions for (p, k, n, n_1) under which RoMP is guaranteed to recover β^* with correct support and small error. We consider the following sub-Gaussian design model. Recall that a random variable Z is sub-Gaussian with parameter σ if $\mathbb{E}[\exp(tZ)] \leq \exp(t^2 \sigma^2 / 2)$ for all real t.

Definition 7 (Sub-Gaussian design). Suppose the entries of $X^{\mathcal{A}}$ are i.i.d. zero-mean sub-Gaussian variables with parameter $\frac{1}{\sqrt{n}}$ and variance $\frac{1}{n}$, and the entries of the additive noise are i.i.d. zero-mean sub-Gaussian variables with parameter $\frac{\sigma_e}{\sqrt{n}}$ and with variance $\frac{\sigma_e^2}{n}$.

Note that this general model covers the case of Gaussian, symmetric Bernoulli, and any other distributions with bounded support.

A. Guarantees for the Distributed Corruption Model

The following theorem characterizes the performance of RoMP, and shows that it can recover the correct support even when the number of outliers scales with n. In particular, this shows RoMP can tolerate an $O(1/\sqrt{k})$ fraction of distributed outliers. Recall that with high probability means with probability at least $1 - p^{-2}$.

Theorem 8. *Under the Sub-Gaussian design model and the distributed corruption model, the following hold with high probability.*

(1) The output of RoMP satisfies the following ℓ_2 error bound:

$$\left\| \hat{\beta} - \beta^* \right\|_2 \ \lesssim \ \|\beta^*\|_2 \sqrt{1 + \frac{\sigma_e^2}{\|\beta^*\|_2^2}} \left(\sqrt{\frac{k \log p}{n}} + \frac{n_1 \sqrt{k} \log p}{n} \right).$$

(2) If the nonzero entries of β^* satisfy $|\beta_j^*|^2 \geq (\|\beta^*\|_2^2/n) \log p \left(1 + \sigma_e^2 / \|\beta^*\|_2^2\right)$, then RoMP correctly identifies the nonzero entries of β^* provided

$$n \ \gtrsim \ k \log p \cdot \left(1 + \sigma_e^2 / \|\beta^*\|_2^2\right), \ and$$

$$\frac{n_1}{n} \ \lesssim \ 1 / \left(\sqrt{k \left(1 + \sigma_e^2 / \|\beta^*\|_2^2\right)} \log p \right).$$

The proof of the theorem is given in Section IX. A few remarks are in order.

1) We emphasize that **knowledge of the exact number of outliers is not needed** – n_1 can be any **upper bound** of the number of outliers, because by definition the adversary can change n_1 entries in each column arbitrarily, and changing less than n_1 of them is of course allowed. The theorem holds even if there are less than n_1 outliers. Of course, this would result in sub-optimal bounds in the estimation due to over-conservativeness. In practice, cross-validation could be quite useful here.

2) We wish to note that essentially all robust statistical procedures we are aware of have the same character noted above. This is true even for the simplest algorithms for robustly estimating the mean. If an upper bound is known on the fraction of corrupted points, one computes the analogous trimmed mean. Otherwise, one can simply compute the median, and the result will have controlled error (but will be suboptimal) as long as the number of corrupted points is less than 50% – something which, as in our case, and every case, is always impossible to know simply from the data.

3) In a similar spirit, the requirement of the knowledge of k can also be relaxed. For example, if we use some $k' > k$ instead of k, then under the theorem continues to hold in the sense that RoMP identifies a superset (with size k') of the support of β^*, and the ℓ_2 error bound holds with k replaced by k'. Standard procedures of estimating the sparsity level (e.g. cross-validation) can also be applied in our setting.

4) Also note that the term $\sigma_e^2 / \|\beta^*\|_2^2$ has a natural interpretation of signal to noise ratio.

B. Guarantee for the Row Corruption Model

It follows directly from Theorem 8 that RoMP can handle $n_1/2$ corrupted rows under the same condition. In fact, a slightly stronger results holds: RoMP can handle n_1 corrupted rows. This is the content of the follow theorem, with its proof given in Section IX.

Corollary 9. *Under the Sub-Gaussian design model and the row corruption model with at most n_1 corrupted rows, the conclusions of Theorem 8 holds.*

Therefore, in particular, our algorithm is orderwise stronger than the Brute Force algorithm, in terms of the number of outliers it can tolerate while still correctly identifying the support.

VIII. EXPERIMENTS

In this section, we report some simulation results for the performance of RoMP (Algorithm 1) on synthetic data. The performance is measured in terms of support recovery (the number of non-zero locations of β^* that are correctly identified), and also relative ℓ_2-error ($\|\hat{\beta} - \beta^*\|_2 / \|\beta^*\|_2$). The authentic data are generated under the sub-Gaussian Design model using Gaussian distribution with $p = 4000, n = 1600, k = 10$ and $\sigma_e = 2$, with the non-zero elements of β^* being randomly assigned to ± 1.

For comparison, we also apply standard Lasso and JP [20], [22] to the same data. For these algorithms, we search for the values of the tradeoff parameters λ, γ that yield the smallest ℓ_2-errors, and then estimate the support using the location of the largest k entries of $\hat{\beta}$. It is also interesting to ask whether a simple modification of JP would perform well. While not analyzed in any of the papers that discuss JP, we consider JP with two different pre-processing procedures, both of which aim to detect and correct the corrupted entries in X directly. The first one, dubbed JP-fill, finds the set E of the largest $\frac{n_1}{n}$ portion of the entries of X, and then scales them to have unit magnitude. The second one, dubbed JP-row, discards the n_1 rows of X that contain the most entries in E.

The corrupted rows $(X^{\mathcal{O}}, y^{\mathcal{O}})$ are generated by the following procedure:

Let

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^{p-k} : \|\theta\|_1 \le \|\beta^*\|_1} \|y^{\mathcal{A}} - X^{\mathcal{A}}_{(\Lambda^*)^\top} \theta\|_2.$$

Set $X^{\mathcal{O}}_{\Lambda^*} = \frac{3}{\sqrt{n}} A$, where A is a random ± 1 matrix of dimension $n_1 \times k$, and $y^{\mathcal{O}} = X^{\mathcal{O}}_{\Lambda^*}(-\beta^*)$. For $i = 1, \ldots, n_1$, further set

$$X^{\mathcal{O}}_{i,(\Lambda^*)^c} = \left(y^{\mathcal{O}}_i / (B_i^\top \theta^*) \right) \cdot B_i^\top,$$

where B_i is a $(n-k)$-vector with i.i.d. standard Gaussian entries.

The results are shown in Figures 2. It can be observed that RoMP performs better than Lasso and JP for both metrics, especially when the number of outliers is large. The ℓ_2-errors of Lasso and JP flatten out because it returns a near-zero solution. The pre-processing procedures do not significantly improve performance of JP, which highlights the difficulty of performing outlier detection in high dimensions.
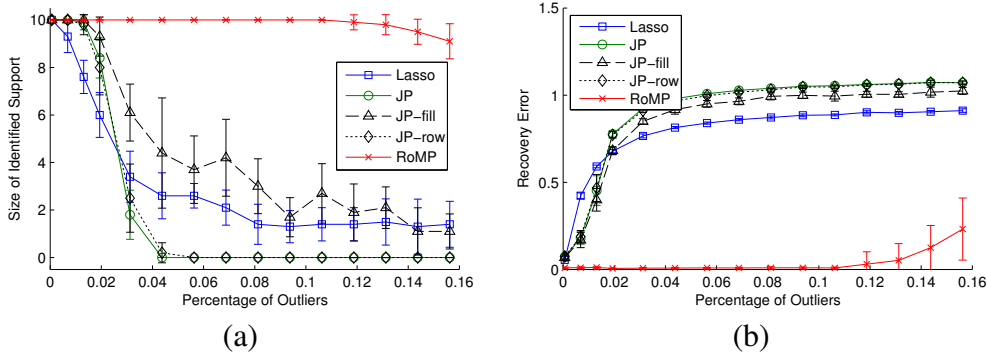

Figure 2: Panel (a) shows the relative support recovery for different methods, while Panel (b) shows the ℓ_2 recovery errors. In both panels, the error bars correspond to one standard deviation. RoMP signficantly outperforms all other methods under both metrics. We note that the ℓ_2 error of the other methods flattens out and error bars shrink to zero, because after about 4% outliers, they return a near-zero solution.

IX. PROOFS

In this section, we turn to the proofs of Theorem 6, Theorem 8, and Corollary 9, with some technical aspects of the proofs deferred to the appendix.

A. Proof of Theorem 2

For simplicity we assume $\Lambda^* = \{1, \ldots, k\}$, $\mathcal{A} = \{1, \ldots, n\}$, and $\mathcal{O} = \{n+1, \ldots, n+n_1\}$. We will show that the adversary choose $y^{\mathcal{O}}$ and $X^{\mathcal{O}}$ in such a way that any "correct" solution of the form $(\theta, \mathcal{S}, \Lambda^*)$ (i.e., with the correct support Λ^*) is not optimal because an alternative solution $(\hat{\theta}, \hat{\mathcal{S}}, \hat{\Lambda})$ with $\hat{\theta} = [1, \ldots, 1]^\top$, $\hat{\mathcal{S}} = \{n_1+1, \ldots, n+n_1\}$, $\hat{\Lambda} = \{2, \ldots, k, k+1\}$ has smaller objective value.

Now for the details. The adversary chooses $\left(y^{\mathcal{O}} \right)_i = \sqrt{k}$ for all i, $X^{\mathcal{O}}_{\Lambda^*} = 0$, and $X^{\mathcal{O}}_{k+1} = y^{\mathcal{O}}$, hence $y^{\mathcal{O}} - X^{\mathcal{O}}_{\hat{\Lambda}} \hat{\theta} = 0$. To compute the objective values of the "correct" solution and the alternative solution, we need a simple technical lemma, which follows from standard results for the norms of random Gaussian matrix. The proof is given in the appendix.

Lemma 10. *If $n \gtrsim k^3 \log p$, we have*

$$\left\| e + X^{\mathcal{A}}_{\Lambda^*} \delta \right\|^2_2 \geq \left(1 - \frac{1}{k} \right) \sigma_e^2, \forall \delta \in \mathbb{R}^k$$

$$\left\| e^{\hat{\mathcal{S}}/\mathcal{O}} + X^{\hat{\mathcal{S}}/\mathcal{O}}_1 - X^{\hat{\mathcal{S}}/\mathcal{O}}_{k+1} \right\|^2_2 \leq \left(1 + \frac{1}{k} \right) \left(1 - \frac{n_1}{n} \right) (\sigma_e^2 + 2)$$

with high probability.

Using the above lemma, we can upper-bound the objective value of the alternative solution:

$$
\begin{aligned}
\left\| y^{\hat{\mathcal{S}}} - X_{\hat{\Lambda}}^{\hat{\mathcal{S}}} \hat{\theta} \right\|_2^2
&= \left\| y^{\mathcal{O}} - X_{\hat{\Lambda}}^{\mathcal{O}} \hat{\theta} \right\|_2^2 + \left\| y^{\hat{\mathcal{S}}/\mathcal{O}} - X_{\hat{\Lambda}}^{\hat{\mathcal{S}}/\mathcal{O}} \hat{\theta} \right\|_2^2 \\
&= 0 + \left\| y^{\hat{\mathcal{S}}/\mathcal{O}} - X_{\Lambda^*}^{\hat{\mathcal{S}}/\mathcal{O}} \beta_{\Lambda^*}^* + X_1^{\hat{\mathcal{S}}/\mathcal{O}} - X_{k+1}^{\hat{\mathcal{S}}/\mathcal{O}} \right\|_2^2 {}_2 \\
&= \left\| e_{\hat{\mathcal{S}}/\mathcal{O}} + X_1^{\hat{\mathcal{S}}/\mathcal{O}} - X_{k+1}^{\hat{\mathcal{S}}/\mathcal{O}} \right\|_2^2 \\
&\leq \left(1 + \frac{1}{k}\right)\left(1 - \frac{n_1}{n}\right)(\sigma_e^2 + 2).
\end{aligned}
\tag{6}
$$

To lower-bound the objective value of solutions of the form $(\theta, \mathcal{S}, \Lambda^*)$, we distinguish two cases. If $\mathcal{S} \cap \mathcal{O} \neq \phi$, then the objective value is

$$
\begin{aligned}
\left\| y^{\mathcal{S}} - X_{\Lambda^*}^{\mathcal{S}} \theta \right\|_2^2
&\geq \left\| y^{\mathcal{S} \cap \mathcal{O}} - X_{\Lambda^*}^{\mathcal{S} \cap \mathcal{O}} \theta \right\|_2^2 \\
&= \left\| y^{\mathcal{S} \cap \mathcal{O}} \right\|_2^2 \\
&\geq k
\end{aligned}
\tag{7}
$$

If $\mathcal{S} \cap \mathcal{O} = \phi$, we have $\mathcal{S} = \mathcal{A}$ and thus

$$
\begin{aligned}
\left\| y^{\mathcal{S}} - X_{\Lambda^*}^{\mathcal{S}} \theta \right\|_2^2
&= \left\| y^{\mathcal{A}} - X_{\Lambda^*}^{\mathcal{A}} \theta \right\|_2^2 \\
&\quad \left\| X_{\Lambda^*}^{\mathcal{A}} \beta_{\Lambda^*}^* + e - X_{\Lambda^*}^{\mathcal{A}} \theta \right\|_2^2 \\
&= \left\| e + X_{\Lambda^*}^{\mathcal{A}} (\beta^* - \theta) \right\|_2^2 \\
&\geq \left(1 - \frac{1}{k}\right)\sigma_e^2
\end{aligned}
\tag{8}
$$

where we use the lemma. When $n_1 > \frac{3n}{k+1}$ and $\sigma_e^2 = k$, we have $\min\left\{k, \left(1 - \frac{1}{k}\right)\sigma_e^2\right\} > \left(1 + \frac{1}{k}\right)\left(1 - \frac{n_1}{n}\right)(\sigma_e^2 + 2)$. Combining (6) (7) and (8) concludes the proof.

B. Proof of Theorem 3

We prove Theorem 3 in this section. We need two technical lemmas. The first lemma bounds the maximum of independent sub-Gaussian random variables. The proof follows from the definition of sub-Gaussianity and Chernoff bound, and is given in the appendix.

Lemma 11. *Suppose $Z_1, \ldots, Z_m$ are m independent sub-Gaussian random variables with parameter σ. Then we have $\max_{i=1,\ldots,m} |Z_i| \leq 4\sigma\sqrt{\log m + \log p}$. with high probability.*

The second lemma is a standard concentration result for the sum of squares of independent sub-Gaussian random variables. It follows directly from Eq. (72) in [23].

Lemma 12. *Let $Y_1, \ldots, Y_n$ be n i.i.d. zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$. Then we have*

$$
|\sum_{i=1}^{n} Y_i^2 - 1| \leq c_1 \sqrt{\frac{\log p}{n}}
$$

with high probability for some absolute constant c_1. Moreover, if $Z_1, \ldots, Z_n$ are also i.i.d. zero-mean sub-Gaussian random variables with parameter $\frac{1}{\sqrt{n}}$ and variance at most $\frac{1}{n}$, and independent of $Y_1, \ldots, Y_n$, then

$$
|\sum_{i=1}^{n} Y_i Z_i| \leq c_2 \sqrt{\frac{\log p}{n}}
$$

with high probability for some absolute constant c_2.

Remark 13. When the above inequality holds, we write $\sum_{i=1}^{n} Y_i^2 \approx 1 \pm \sqrt{\frac{\log p}{n}}$ and $\sum_{i=1}^{n} Y_i Z_i \approx \pm \sqrt{\frac{\log p}{n}}$. w.h.p.

Now consider the trimmed inner product $h(j)$ between the jth column of X and y. Let $\mathcal{A}_j$ is the set of index i such that X_{ij} and y_i are both not corrupted. By assumption $|\mathcal{A}_j| \geq n$. By putting $|\mathcal{A}_j| - n$ clean indices in $\mathcal{A}_j^c$, we may assume $|\mathcal{A}_j| = n$ without loss of generality. By prescription of Algorithm 2, we can write $h(j)$ as

$$h(j) = \sum_{i \in \mathcal{A}_j} X_{ij} y_i - \sum_{\substack{i \in \text{trimmed} \\ \text{inliers}}} X_{ij} y_i + \sum_{\substack{i \in \text{remaining} \\ \text{outliers}}} X_{ij} y_i.$$

We estimate each term in the above sum.

1) Observe that

$$\sum_{i \in \mathcal{A}_j} X_{ij} y_i = \sum_{i \in \mathcal{A}_j} X_{ij} \left(\sum_{k=1}^{p} X_{ik} \beta_k^* + e \right) = \sum_{i \in \mathcal{A}_j} X_{ij}^2 \beta_j^* + \sum_{i \in \mathcal{A}_j} X_{ij} \left(\sum_{k \neq j} X_{ik} \beta_k^* + e \right).$$

(a) Because the points in $\mathcal{A}_j$ obeys the Sub-Gaussian model, Lemma 12 gives $\sum_{i \in \mathcal{A}_j} X_{ij}^2 \beta_j^* \approx \beta_j^* \left(1 \pm \sqrt{\frac{1}{n} \log p} \right)$ w.h.p.

(b) On the other hand, because X_{ik} and X_{ij} are independent when $k \neq j$, and $Z_i \triangleq \sum_{k \neq j} X_{ik} \beta_k^* + e$ are i.i.d. sub-Gaussian with parameter and standard deviation at most $\sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right) / n}$, we apply Lemma 12 to obtain $\sum_{i \in \mathcal{A}_j} X_{ij} Z_i \approx \pm \frac{1}{\sqrt{n}} \sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right) \log p}$ w.h.p.

2) Again due to independence and sub-Gaussianity of points in $\mathcal{A}_j$, Lemma 11 gives $\max_{i \in \mathcal{A}_j} |X_{ij}| \lesssim \sqrt{(\log p)/n}$ w.h.p. and $\max_{i \in \mathcal{A}_j} |y_i| \lesssim \sqrt{(\log p/n) \left(\|\beta^*\|_2^2 + \sigma_e^2 \right)}$ w.h.p. It follows that w.h.p.

$$\left| \sum_{\substack{i \in \text{trimmed} \\ \text{inliers}}} X_{ij} y_i \right| \leq n_1 \left(\max_{i \in \mathcal{A}} |X_{ij}| \right) \left(\max_{i \in \mathcal{A}} |y_i| \right) \lesssim n_1 \cdot \sqrt{\frac{\log p}{n}} \cdot \sqrt{\frac{\log p}{n} \left(\|\beta^*\|_2^2 + \sigma_e^2 \right)}.$$

3) By prescription of the trimming procedure, either all outliers are trimmed, or the remaining outliers are no larger than the trimmed inliers. It follows from the last equation that w.h.p.

$$\left| \sum_{\substack{i \in \text{remaining} \\ \text{outliers}}} X_{ij} y_i \right| \leq \sum_{\substack{i \in \text{remaining} \\ \text{outliers}}} |X_{ij} y_i| \leq \sum_{\substack{i \in \text{trimmed} \\ \text{inliers}}} |X_{ij} y_i| \lesssim n_1 \frac{\log p}{n} \cdot \sqrt{\|\beta^*\|_2^2 + \sigma_e^2}.$$

Combining pieces, we have for all $j = 1, \ldots, p$,

$$\left| h(j) - \beta_j^* \right| \lesssim |\beta_j^*| \sqrt{\frac{2}{n} \log p} + \frac{1}{\sqrt{n}} \sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right) \log p} + n_1 \cdot \frac{\log p}{n} \sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right)}. \tag{9}$$

If RoMP correctly picks an index j in the true support Λ^*, then the error in estimating $\hat{\beta}_j$ is bounded by the expression above. If RoMP picks some incorrect index j not in Λ^*, then the difference between the corresponding $\hat{\beta}_j$ and the true $\beta_{j'}^*$ that should have been picked is still bounded by the expression above (up to constant factors). Therefore, we have

$$\left\| \hat{\beta} - \beta^* \right\|_2^2 \lesssim \sum_{j \in \Lambda^*} \left[|\beta_j^*| \sqrt{\frac{2}{n} \log p} + \frac{1}{\sqrt{n}} \sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right) \log p} + n_1 \frac{\log p}{n} \sqrt{\left(\|\beta^*\|_2^2 + \sigma_e^2 \right)} \right]^2.$$

The first part of the theorem then follows after straightforward algebra manipulation. On the other hand, RoMP picks the correct support as long as $|h(j)| > |h(j')|$ for all $j \in \Lambda^*, j' \in (\Lambda^*)^c$. In view of Eq.(9), we require

$$
n \gtrsim \max_j \left(\frac{\|\beta^*\|_2^2}{\beta_j^2} \right) \cdot \log p \cdot \left(1 + \sigma_e^2 / \|\beta^*\|_2^2 \right)
$$

$$
\frac{n_1}{n} \lesssim \frac{1}{\sqrt{\max_j \left(\frac{\|\beta^*\|_2^2}{\beta_j^2} \right) \cdot \left(1 + \sigma_e^2 / \|\beta^*\|_2^2 \right) \log p}}
$$

One verifies that the above inequalities are satisfied under the conditions in the second part of the theorem.

C. Proof of Corollary 1

A careful examination of the proof of Theorem 2 in the last section shows that, when there are n_1 corrupted rows, the set $\mathcal{A}_j$ still has cardinality at least n, and the proof thus holds under the row corruption model.

X. Conclusion

Adversarial corruption seems to be significantly more difficult than corruption independent from the original data, and moreover, corruption in X as well as y appears more challenging than corruption only in y. To the best of our knowledge, no prior existing algorithms have provable performance in this setting, or in the more difficult yet setting of distributed corruption. This paper provides the first results for both these settings. Our results outperform Justice Pursuit, as well as the exponential time Brute Force algorithm; more generally we show that no convex optimization based approach improve on the results we provide. Generalizing our results to obtain a sequential OMP-like algorithm, and on the other side, understanding converse results, are important next steps.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[2] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 2011.

[3] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[4] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[5] E.J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.

[6] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[7] V. Chandrasekaran, S. Sanghavi, S. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[8] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1999.

[9] Y. Chen and C. Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high dimensional results. *arXiv preprint arXiv:1206.0823*, 2012.

[10] Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, 2013.

[11] Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust matrix completion with corrupted columns. *Submitted. Arxiv Preprint arXiv:1102.2254v1*, 2011.

[12] M.A. Davenport and M.B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *Information Theory, IEEE Transactions on*, 56(9):4395–4401, 2010.

[13] D. L. Donoho. Breakdown properties of multivariate location estimators, qualifying paper, Harvard University, 1982.

[14] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[15] J.J. Fuchs. An inverse problem approach to robust regression. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, volume 4, pages 1809–1812. IEEE, 1999.

[16] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. Wiley, 1986.

[17] M.A. Herman and T. Strohmer. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, 2010.

[18] Peter Huber. *Robust Statistics*. Wiley, New York, 1981.

[19] V. Kekatos and G.B. Giannakis. From sparse signals to sparse residuals for robust sensing. *Signal Processing, IEEE Transactions on*, 59(7):3355–3368, 2011.

[20] J.N. Laska, M.A. Davenport, and R.G. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 1556–1560. IEEE, 2009.

[21] G. Lerman, M. McCoy, J.A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *Arxiv preprint arXiv:1202.4044*, 2012.

[22] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Arxiv preprint arXiv:1104.1041*, 2011.

[23] P.L. Loh and M.J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[24] R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust statistics*. Wiley, 2006.

[25] N.H. Nguyen, T. Tran, et al. Exact recoverability from dense corrupted observations via l_1 minimization. *Arxiv preprint arXiv:1102.1227*, 2011.

[26] M. Rosenbaum and A.B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[27] M. Rosenbaum and A.B. Tsybakov. Improved matrix uncertainty selector. *arXiv preprint arXiv:1112.4413*, 2011.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[29] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

[30] M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Information Theory, IEEE Transactions on*, 55(12):5728–5741, 2009.

[31] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[32] J. Wright and Y. Ma. Dense error correction via-minimization. *Information Theory, IEEE Transactions on*, 56(7):3540–3560, 2010.

[33] H. Xu, C. Caramanis, and S. Mannor. Outlier-Robust PCA: The High Dimensional Case. *IEEE Transactions on Information Theory*, 59(1):546–572, 2013.

[34] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

[35] Y. Yu, O. Aslan, and D. Schuurmans. A polynomial-time form of robust regression. In *Advances in Neural Information Processing Systems 25*, pages 2492–2500, 2012.

[36] H. Zhu, G. Leus, and G.B. Giannakis. Sparsity-cognizant total least-squares for perturbed compressive sampling. *Signal Processing, IEEE Transactions on*, 59(5):2002–2016, 2011.

APPENDIX

Let $\theta' = \left[\sigma_e | \delta^\top \right]^\top$. We can write $\left\| e + X_{\Lambda^*}^{\mathcal{A}} \delta \right\|_2^2 = \left\| Z_1 \theta' \right\|_2^2$ with $Z_1 \triangleq \left[\frac{1}{\sigma_e} e | X_{\Lambda^*}^{\mathcal{A}} \right]$. Note that Z_1 is an $n \times (k+1)$ matrix with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries, whose smallest singular value can be bounded using standard results. For example, using Lemma 5.1 in [1] with $\Phi(\omega) = Z_1$, $N = k+1$, $T = \{1, \ldots, N\}$, $\delta = \frac{1}{3k}$ and $c_0(\delta/2) = 1/288k^2$, we have

$$\|Z_1 \theta'\|_2^2 \leq \left(1 + \frac{1}{3k} \right)^2 \|\theta'\|_2^2 \leq \left(1 + \frac{1}{k} \right) \sigma_e^2, \forall \delta$$

with probability at least

$$1 - 2e^{\frac{1}{288k^2} n - (k+1) \ln(36k)} \geq 1 - 2p^{-3}$$

provided $n \geq 576(k+1)^3 \ln(36p)$. This proves the first inequality.

Let $\hat{Z} = \max_i Z_i$. By definition of sub-Gaussianity, we have

$$\begin{aligned}
\mathbb{E}\left[e^{t\hat{Z}/\sigma} \right] &= \mathbb{E}\left[\max_i e^{tZ_i/\sigma} \right] \\
&\leq \sum_i \mathbb{E}\left[e^{tZ_i/\sigma} \right] \\
&\leq m e^{t^2/2} \\
&= e^{t^2/2 + \log m}
\end{aligned}$$

It follows from Markov Inequality that

$$
\begin{aligned}
P(\hat{Z} \geq \sigma t) &= P(e^{t\hat{Z}/\sigma} \geq e^{t^2}) \\
&\leq e^{-t^2} \mathbb{E}\left[e^{t\hat{Z}/\sigma}\right] \\
&\leq e^{-t^2 + t^2/2 + \log m} \\
&= e^{-\frac{1}{2}t^2 + \log m}.
\end{aligned}
$$

By symmetry we have

$$
P(\min_i Z_i \leq -\sigma t) \leq e^{-\frac{1}{2}t^2 + \log m},
$$

so a union bound gives

$$
\begin{aligned}
P(\max_i |Z_i| \geq \sigma t) &\leq P(\max_i Z_i \geq \sigma t) + P(\min_i Z_i \leq -\sigma t) \\
&\leq 2e^{-\frac{1}{2}t^2 + \log m}.
\end{aligned}
$$

Taking $t = 4\sqrt{\log m + \log p}$ yields the result.